(19)中华人民共和国国家知识产权局





(12)发明专利申请

(10)申请公布号 CN 109286576 A
(43)申请公布日 2019.01.29

(21)申请号 201811176875.1

(22)申请日 2018.10.10

(71)申请人 北京理工大学
地址 100081 北京市海淀区中关村南大街5号

(72)发明人 沈蒙 张晋鹏 祝烈煌 徐恪

(74)专利代理机构 北京理工正阳知识产权代理事务所(普通合伙) 11639
代理人 唐华

(51)Int.Cl.
H04L 12/851(2013.01)
H04L 12/24(2006.01)
G06K 9/62(2006.01)
H04L 29/06(2006.01)

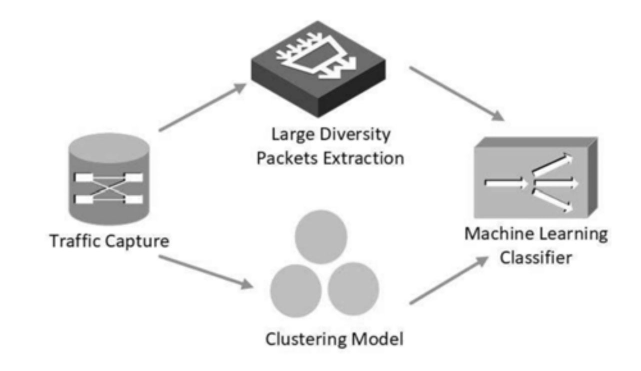
权利要求书2页 说明书6页 附图2页

(54)发明名称

一种数据包频度分析的网络代理加密流量特征提取方法

(57)摘要

本发明涉及一种数据包频度分析的网络代理加密流量特征提取方法,属于机器学习以及网络服务安全领域。包括如下步骤:步骤1、基于数据包频度分析结果抽取数据包;步骤2、数据包长度-时间戳之差聚类,生成聚类结果;步骤3、计算最优类簇数量;步骤4、计算加密流量特征。所述方法引入了基于词频逆文档频率的区分度较大数据包,比直接使用所有数据包更有明显的区分作用;能作用于任何机器学习分类算法上,分类准确率高;引入了数据包长度和时间戳之差聚类,可进一步提升URL不同页面元素相同的网页的分类效果;与现有的加密网络流量分类和识别方法相比具有更高的准确性。



CN 109286576 A

1.一种数据包频度分析的网络代理加密流量特征提取方法,其特征在于:包括如下步骤:

步骤1、基于数据包频度分析结果抽取数据包;

其中,抽取数据包具体为:将数据包频度分析结果中区分度大的数据包抽取出来;

其中,区分度大的数据包是指词频-逆文档频率不小于0.00001的数据包;

步骤1又包括如下子步骤:

步骤1.1对捕获到的数据包进行数据包编码,得到编码后数据包;

其中,捕获到的数据包为TCP数据包,用于区分TCP数据包的标志位有[SYN]、[SYN,ACK]、[ACK]、[PSH,ACK]和[FIN,ACK];

其中,[SYN]表示客户端和服务器之间建立TCP连接时的SYN消息,[SYN,ACK]表示客户端和服务器建立连接时的服务器的应答,[ACK]表示收到消息的确认,[PSH,ACK]表示发出消息的同时对收到的消息进行确认,[FIN,ACK]表示通信双方断开连接;

步骤1.2计算步骤1.1输出的编码后数据包的词频$tf_{i,j}$,$tf_{i,j}$代表第i种数据包在第j类网页流量中的比例,遍历i和j,又具体包括如下子步骤:

步骤1.2A统计第j类网页流量中第i种数据包的个数$n_{i,j}$;

步骤1.2B统计第j类网页中的所有数据包个数总和为$\sum_k n_{k,j}$;

步骤1.2C用第i种数据包的个数$n_{i,j}$除以第j类网页的所有数据包个数,即通过(1)计算第i种数据包在第j类网页中的词频$tf_{i,j}$:

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (1)$$

其中,k代表第j类网页中的数据包种数;

步骤1.3计算步骤1.1输出的编码后数据包的逆文档频率;

特定数据包i在网页j流量中的计数为$|\{j:t_i\}\in d_j|$,所有网页流量总数为$|D|$,通过(2)计算第i种数据包的逆文档频率idf_i:

$$idf_i=log\frac{|D|}{1+|\{j:t_i\in d_j\}|} \qquad (2)$$

其中,log是以10为底的对数操作;

步骤1.4根据步骤1.2和步骤1.3计算得到的词频$tf_{i,j}$、逆文档频率idf_i,通过(3)计算第i种数据包在第j类网页中的词频-逆文档频率$TI_{i,j}$:

$TI_{i,j}=tf_{i,j}\times idf_i$　　(3)

步骤1.5根据步骤1.4得到的词频-逆文档频率$TI_{i,j}$,去掉词频-逆文档频率小于0.00001的数据包,选择剩下的数据包用作分类;

步骤2、数据包长度-时间戳之差聚类,生成聚类结果,具体为:

步骤2.1提取网页流量中每条流的第一个上行[PSH,ACK]数据包的长度l_p,所有流的第一个上行[PSH,ACK]数据包长度汇集在一个文件中;

步骤2.2提取每条流的第一个上行[PSH,ACK]数据包的时间戳信息t_u,接着提取每条流的第一个下行[PSH,ACK]数据包的时间戳信息t_d;再将下行[PSH,ACK]数据包的时间戳信息t_d减去上行的时间戳信息t_u的结果作为时间戳之差t,保存所有网络流的时间戳之差;

步骤2.3将每条流中的第一个上行[PSH,ACK]数据包的长度和时间戳之差保存在一个

文件中供聚类使用；

步骤2.4遍历簇数m从2到q_{max}，将步骤2.1提取的数据包长度l_p和时间戳之差t进行聚类，生成聚类结果Cm；

其中，q_{max}代表最大的类簇数量；

$q_{max}=J\times 3$　(4)

其中，J为要分类网页的类数；

其中，聚类采用K-Means方法；

聚类结果，记为Cm＝{$cent_1$，…，$cent_m$}，$cent_m$代表第m个类簇中心的中心值；

其中，每条流中要参与聚类的元素为(l_p，t)，两个聚类点$clup_a$，$clup_b$之间的距离dis($clup_a$，$clup_b$)采用公式(5)计算：

$$\mathrm{dis}(clup_a, clup_b)=\sqrt{\left(l_{p_a}-l_{p_b}\right)^2+(t_a-t_b)^2} \tag{5}$$

步骤3计算最优类簇数量，具体为：

步骤3.1遍历ω基于(6)计算聚类点clup与类簇中心$cent_\omega$的距离和SSE(ω)：

$$\mathrm{SSE}(\omega)=\sum_{r=1}^{P}\sum_{\omega=1}^{m}\|clup_r-cent_\omega\|^2 \tag{6}$$

其中，P代表聚类点clup的个数；m的取值范围为2到q_{max}；

步骤3.2选择步骤3.1计算的最小SSE(ω)对应的类簇中心数量为最优类簇数量，此最小的SSE(ω)记为SSE(ω_{opt})，此最小SSE(ω)对应的最优类簇中心记为Cm(ω_{opt})；

步骤4计算加密流量特征，具体包括如下子步骤：

步骤4.1计算步骤1中提取出来的区分度大的数据包的统计特征值(max，min，mean，…，var)；

步骤4.2计算每条流中上行第一个[PSH，ACK]数据包的大小与时间戳之差形成的二元组与步骤3生成的最优类簇中心Cm(ω_{opt})之间的距离$\left(dis_1,\ldots,dis_{\omega_{opt}}\right)$；

其中，步骤4.1的统计特征值(max，min，mean，…，var)与步骤4.2的二元组与类簇中心的距离$\left(dis_1,\ldots,dis_{\omega_{opt}}\right)$作为加密流F的特征。

2.如权利要求1所述的一种数据包频度分析的网络代理加密流量特征提取方法，其特征在于：步骤1.1中，数据包编码结合标志位、数据包的长度信息及数据包的方向信息进行综合编码；

其中，数据包的方向用U、D表示，U代表上行，D代表下行。

3.如权利要求1所述的一种数据包频度分析的网络代理加密流量特征提取方法，其特征在于：步骤1.2中的i和j均大于1。

4.如权利要求1所述的一种数据包频度分析的网络代理加密流量特征提取方法，其特征在于：步骤2.1中的每条流通过将网页流量中按源端口、目的端口、源IP、目的IP和协议五元组进行划分得到。

一种数据包频度分析的网络代理加密流量特征提取方法

技术领域

[0001]　本发明涉及一种数据包频度分析的网络代理加密流量特征提取方法，尤其涉及一种基于数据包频度与数据包长度和时间戳之差聚类的机器学习Shadowsocks代理的加密流量特征提取方法，旨在为识别Shadowsocks加密过后的网页流量提供区分度大的流量特征，属于机器学习以及网络服务安全领域。

背景技术

[0002]　流量是网络信息传输的载体。Shadowsocks是一种基于SOCKS5的加密代理技术，作用在传输层和应用层之间为用户提供代理服务。本发明所指流量识别技术是对经过Shadowsocks加密过后的流量进行细粒度的分类识别。通过对Shadowsocks流量进行细粒度的分类识别，可以对用户的上网习惯进行分析，也可以及时发现恶意页面的流量实施有效的拦截和屏蔽，保障网络安全。在实际应用中，通过将该类识别功能部署在路由器等网关节点中，可以及早发现并屏蔽恶意页面的流量，确保网络安全。

[0003]　现有的流量识别方法主要包含两大类：明文流量识别和加密流量识别。在明文流量识别中采取的主要技术是深度数据包检测和端口检测。随着加密技术的采用和跳变端口技术的采用，网络通信过程中的数据包被加密，深度数据包检测技术和端口检测技术逐渐失去了效用。现在的研究热点主要集中在加密流量识别中。流量的加密技术主要有两种：SSL/TLS(安全套接层/传输层安全)协议和基于Socks5的加密代理协议。目前针对标准SSL/TLS加密后的流量识别技术研究比较充分，而针对基于Socks5的加密代理流量的识别则不是很充分。Shadowsocks是一种基于Socks5的加密代理技术。

[0004]　在Shadowsocks加密网络流量分类和识别方面，可检索到的关联最大的两项专利为：

[0005]　(1)现有文献提出两种分类Secure Shell(SSH)协议加密后的流量识别方法。研究者用到的数据包分类特征为数据包的大小和数据包的方向。通过对数据包大小和方向向量化表示，作者采用高斯混合模型(Gaussian Mixture Models,GMM)和支持向量机(Support Vector Machines,SVM)对SSH协议加密过后的网络流量进行分类。该种分类方法的识别是粗粒度识别，可以对应用层的不同协议进行识别，如识别HTTP、POP3和SEMULE等不同应用层协议的流量。

[0006]　(2)已有专利提出了一种在背景流量中检测Shadowsocks流量的方法。研究者将总的数据包个数、流出数据包个数、流入数据包个数、传输时间、流入数据包的比例、流出数据包的比例、最大数据包长度、平均数据包长度等信息进行特征提取，将提取好的特征值放入随机森林(Random Forest)分类器中进行分类可以从背景流量中有效识别出Shadowsocks流量，识别的准确率为85％。这种方法只是从背景流量中识别Shadowsocks流量，不能进行进一步的细粒度的流量分类。

[0007]　综上所述，在SSH流量分类领域有对SSH粗粒度的分类，而对使用Shadowsocks加密后的代理流量识别领域，目前仅有从背景流量中识别Shadowsocks流量的方法，还没有对

Shadowsocks加密过后的流量进行细粒度识别的方法。

发明内容

[0008] 本发明的目的在于为识别Shadowsocks加密过后的网页流量提供区分度大的流量特征，进而辅助于Shadowsocks流量细粒度分类，通过对加密后的网页流量进行分类，对用户的行为习惯进行分析以及检测恶意网页流量，应用于Shadowsocks加密过后的网页流量，提出了一种数据包频度分析的网络代理加密流量特征提取方法。

[0009] 所述网络代理加密流量特征提取方法，包括如下步骤：

[0010] 步骤1、基于数据包频度分析结果抽取数据包；

[0011] 其中，抽取数据包具体为：将数据包频度分析结果中区分度大的数据包抽取出来；

[0012] 其中，区分度大的数据包是指词频-逆文档频率不小于0.00001的数据包；

[0013] 步骤1又包括如下子步骤：

[0014] 步骤1.1对捕获到的数据包进行数据包编码，得到编码后数据包；

[0015] 其中，捕获到的数据包为TCP数据包，数据包编码结合标志位、数据包的长度信息及数据包的方向信息进行综合编码；

[0016] 其中，用于区分TCP数据包的标志位有[SYN]、[SYN,ACK]、[ACK]、[PSH,ACK]和[FIN,ACK]；

[0017] 其中，[SYN]表示客户端和服务器之间建立TCP连接时的SYN消息，[SYN,ACK]表示客户端和服务器建立连接时的服务器的应答，[ACK]表示收到消息的确认，[PSH,ACK]表示发出消息的同时对收到的消息进行确认，[FIN,ACK]表示通信双方断开连接；

[0018] 其中，数据包的方向用U、D表示，U代表上行，D代表下行；

[0019] 步骤1.2计算步骤1.1输出的编码后数据包的词频$tf_{i,j}$，$tf_{i,j}$代表第i种数据包在第j类网页流量中的比例，遍历i和j；

[0020] 其中，i和j均大于1；

[0021] 步骤1.2又具体包括如下子步骤：

[0022] 步骤1.2A统计第j类网页流量中第i种数据包的个数$n_{i,j}$；

[0023] 步骤1.2B统计第j类网页中的所有数据包个数总和为$\sum_k n_{k,j}$；

[0024] 步骤1.2C用第i种数据包的个数$n_{i,j}$除以第j类网页的所有数据包个数，即通过(1)计算第i种数据包在第j类网页中的词频$tf_{i,j}$：

[0025] $$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (1)$$

[0026] 其中，k代表第j类网页中的数据包种数；

[0027] 步骤1.3计算步骤1.1输出的编码后数据包的逆文档频率；

[0028] 特定数据包i在网页j流量中的计数为$|\{j:t_i\}\in d_j|$，所有网页流量总数为$|D|$，通过(2)计算第i种数据包的逆文档频率idf_i：

[0029] $$idf_i=log\frac{|D|}{1+|\{j:t_i\epsilon d_j\}|} \qquad (2)$$

[0030] 其中，log是以10为底的进行取对数；

[0031] 步骤1.4根据步骤1.2和步骤1.3计算得到的词频$tf_{i,j}$、逆文档频率idf_i，通过(3)

计算第i种数据包在第j类网页中的词频-逆文档频率$TI_{i,j}$:

[0032] $TI_{i,j}=tf_{i,j}\times idf_i$ (3)

[0033] 步骤1.5根据步骤1.4得到的词频-逆文档频率$TI_{i,j}$,去掉词频-逆文档频率小于0.00001的数据包,选择剩下的数据包用作分类;

[0034] 步骤2、数据包长度-时间戳之差聚类,生成聚类结果;

[0035] 步骤2.1提取网页流量中每条流的第一个上行[PSH,ACK]数据包的长度l_p,所有流的第一个上行[PSH,ACK]数据包长度汇集在一个文件中;

[0036] 其中,每条流通过将网页流量中按源端口、目的端口、源IP、目的IP和协议五元组进行划分得到;

[0037] 步骤2.2提取每条流的第一个上行[PSH,ACK]数据包的时间戳信息t_u,接着提取每条流的第一个下行[PSH,ACK]数据包的时间戳信息t_d;再将下行[PSH,ACK]数据包的时间戳信息t_d减去上行的时间戳信息t_u的结果作为时间戳之差t,保存所有网络流的时间戳之差;

[0038] 步骤2.3将每条流中的第一个上行[PSH,ACK]数据包的长度和时间戳之差保存在一个文件中供聚类使用;

[0039] 步骤2.4遍历簇数m从2到q_{max},将步骤2.1提取的数据包长度l_p和时间戳之差t进行聚类,生成聚类结果Cm;

[0040] 其中,q_{max}代表最大的类簇数量;

[0041] $q_{max}=J\times 3$ (4)

[0042] 其中,J为要分类网页的类数;

[0043] 其中,聚类采用K-Means方法;

[0044] 聚类结果,记为$Cm=\{cent_1,\cdots,cent_m\}$,$cent_m$代表第m个类簇中心的中心值;

[0045] 其中,每条流中要参与聚类的元素为(l_p,t),两个聚类点$clup_a$,$clup_b$之间的距离$dis(clup_a,clup_b)$采用公式(5)计算:

[0046] $$\mathrm{dis}(clup_a,clup_b)=\sqrt{\left(l_{p_a}-l_{p_b}\right)^2+(t_a-t_b)^2} \quad (5)$$

[0047] 步骤3计算最优类簇数量,具体为:

[0048] 步骤3.1遍历ω基于(6)计算聚类点clup与类簇中心$cent_\omega$的距离和SSE(ω):

[0049] $$\mathrm{SSE}(\omega)=\sum_{r=1}^{P}\sum_{\omega=1}^{m}\|clup_r-cent_\omega\|^2 \quad (6)$$

[0050] 其中,P代表聚类点clup的个数;m的取值范围为2到q_{max};

[0051] 步骤3.2选择步骤3.1计算的最小SSE(ω)对应的类簇中心数量为最优类簇数量,此最小的SSE(ω)记为$SSE(\omega_{opt})$,此最小SSE(ω)对应的最优类簇中心记为$Cm(\omega_{opt})$;

[0052] 步骤4计算加密流量特征,具体包括如下子步骤:

[0053] 步骤4.1计算步骤1中提取出来的区分度大的数据包的统计特征值(max,min,mean,…,var);

[0054] 步骤4.2计算每条流中上行第一个[PSH,ACK]数据包的大小与时间戳之差形成的二元组与步骤3生成的最优类簇中心$Cm(\omega_{opt})$之间的距离$\left(dis_1,\ldots,dis_{\omega_{opt}}\right)$;

[0055] 至此,步骤4.1的统计特征值(max,min,mean,…,var)与步骤4.2的二元组与类簇

中心的距离$\left(dis_1, \ldots, dis_{\omega_{opt}}\right)$作为加密流F的特征。

[0056]　有益效果

[0057]　本发明提出了一种数据包频度分析的网络代理加密流量特征提取方法，与现有网络代理加密流量特征提取方法相比，具有如下有益效果：

[0058]　(1)本发明适用于Shadowsocks代理加密过后的网络流量进行分类；

[0059]　(2)本发明引入了基于词频逆文档频率的区分度较大数据包特征提取技术，用该种技术提取出来的数据包比直接使用所有数据包更有明显的区分作用；

[0060]　(3)本发明引入了的基于词频逆文档频率的区分度较大数据包特征提取技术，这种方法提取出来的数据包能作用于任何机器学习分类算法上，对分类准确率的提升有较大贡献；

[0061]　(3)本发明引入了数据包长度和时间戳之差聚类，可进一步提升URL不同页面元素相同的网页的分类效果；

[0062]　(4)本发明通过大量实验数据实验证明，与现有的加密网络流量分类和识别方法相比具有更高的准确性。

附图说明

[0063]　图1为本发明一种数据包频度分析的网络代理加密流量特征提取方法的整体流程图；

[0064]　图2为本发明一种数据包频度分析的网络代理加密流量特征提取方法步骤1中的词频-逆文档频率区分度大的数据包提取示意图；

[0065]　图3为本发明一种数据包频度分析的网络代理加密流量特征提取方法步骤2中的数据包长度和时间戳之差聚类结果示意图。

具体实施方式

[0066]　下面结合附图和实施例，更具体地说明本发明“基于数据包频度分析的网络代理加密流量特征提取方法”的过程，并阐述其优点。应当指出，本发明的实施并不局限于下面的实施例，对本发明所做的任何形式上的变通或改变将落入本发明保护范围。

[0067]　实施例1

[0068]　本实施例是基于本发明的步骤1到步骤4进行的完整的Shadowsocks加密代理流量特征提取仿真，整体流程图如图1所示，通过区分度大的数据包提取技术和聚类结果共同作用生成的网络流量特征用于加密代理流量分类。

[0069]　首先进行区分度大的数据包抽取，具体流程如图2所示。假设捕获到的某条数据流表示为F＝(p_1,..,p_n)，其中p_i代表第i个数据包。数据包p_i包含的信息包含了三部分数据包方向、数据包大小以及数据包的标志信息，如果数据包p_i为从客户端发往服务器的长度为54的SYN数据包，则该数据包编码为U_54_SYN，代表从客户端发往服务器的长度为54的SYN包；对所有数据包进行上述编码。

[0070]　编码之后计算每种数据包在不同网页流量中的出现的频率。数据包编码为U_54_SYN，U_66_SYNACK，U_54_ACK，U_77_PSHACK，U_671_PSHACK，D_54_ACK，U_1354_ACK，D_54_

FINACK在www.google.com页面流量中出现的频率为0.01785,0.01785,0.03571,0.0714,0.0714,0.3571,0.0714,0.03571。

[0071] 接着计算包含不同数据包的www.google.com页面流量在所有页面流量中的逆文档频率,所有流量条数为5000,其中包含U_54_SYN的流量条数为4500,则U_54_SYN的逆文档频率计算为$\log\frac{5000}{4998+1}=0.0000869$。用每种数据包的频率乘以包含该数据包的网页流量的逆文档频率,即为该种数据包在该网页中的词频-逆文档频率TI,如在本实例中U_54_SYN的数据包在google.com中的文档-逆文档频率为0.01785×0.0000869=0.0008168。google.com所包含的数据包词频-逆文档频率计算结果如表1所示。

[0072] 表1 www.google.com网页流量中数据包词频-逆文档频率计算结果

[0073]

数据包	词频	逆文频率	词频-逆文档频率
U_54_SYN	0.01785	0.0000869	0.00000155
U_66_SYNACK	0.01785	0.0000869	0.00000155
U_54_ACK	0.35714	0.0000869	0.00003102
U_77_PSHACK	0.07142	0.0086853	0.00062030
U_671_PSHACK	0.07142	0.0086853	0.00062030
D_54_ACK	0.03571	0.0086853	0.00031015
D_1354_ACK	0.07142	0.0086853	0.00062030
D_54_FINACK	0.03571	0.0086853	0.00031015

[0074] 根据表1可知U_54_SYN,U_66_SYNACK的数据包区分效果不好,在进行特征值计算时主动滤除这些数据包。

[0075] 进行聚类时首先提取每条流中的第一个[PSH,ACK]数据包的大小,然后提取上行第一个[PSH,ACK]数据包和下行第一个[PSH,ACK]数据包的时间戳之差,然后将时间戳之差扩大1000倍,将数据包大小和时间戳之差组成的二元组信息进行聚类。聚类好的类簇中心为:[(97.57143,732.809),(107.7105,143.8095),…,(1354,702)],聚类结果如图3所示。

[0076] 加密流量特征值计算,将去除区分度低的数据包后的其它数据包计算统计特征值,计算该条流的聚类点(l_p,t)与上述类簇中心的距离组成的向量组作为Shadowsocks加密代理流量的分类特征。

[0077] 实施例2

[0078] 本实施例是将本发明所述方法与其它流量分类算法进行对比,以验证本发明的优势及有效性。将本发明所述的基于词频分析的流量特征提取方法(TF-IDF)与传统机器学习算法最近邻算法(k-NN)、支持向量机(SVM)、随机森林(RANF)结合构建的网络流量分类器的效果要优于不使用直接使用这些分类器进行分类的结果。使用同一流量数据集对网页流量进行分类,不同方法的对比结果如表2所示:

[0079] 表2不同方法分类准确率对比

[0080]

分类算法	k-NN	k-NN_T	SVM	SVM_T	RANF	RANF_T
准确率	67.51%	72.85%	63.62%	72.81%	71.04%	76.16%

[0081] 从表2可以看出使用词频分析过后的抽取的区分度大的数据包无论用于哪种流量分类算法都能提高分类器的准确率，使用随机森林分类器的效果最好。我们将引入聚类模型之后的分类模型记为RFTC，与目前先进的流量分类算法的分类结果对比如表3所示：

[0082] 表3与先进的流量分类模型分类效果对比

[0083]

分类方法	DDTW	APPS	RFTC
精确率	56.71%	71.83%	79.52%
召回率	54.08%	71.04%	79.38%
准确率	54.10%	71.00%	79.39%

[0084]

[0085] 从表3可以看出，本发明与现有的流量分类方法相比，具有明显优势，精确率、召回率及准确率都高于其他两种分类算法。本发明对使用代理之后的流量可以提取良好的流量特征，助力于流量精细化分类检测，能够提高分类准确率，可以投入实际应用中。

[0086] 虽然本文结合附图实例描述了本专利的实施方式，但是对于本领域技术人员来说，在不脱离本专利原理的前提下，还可以做出若干改进，这些也是为属于本专利的保护范围。

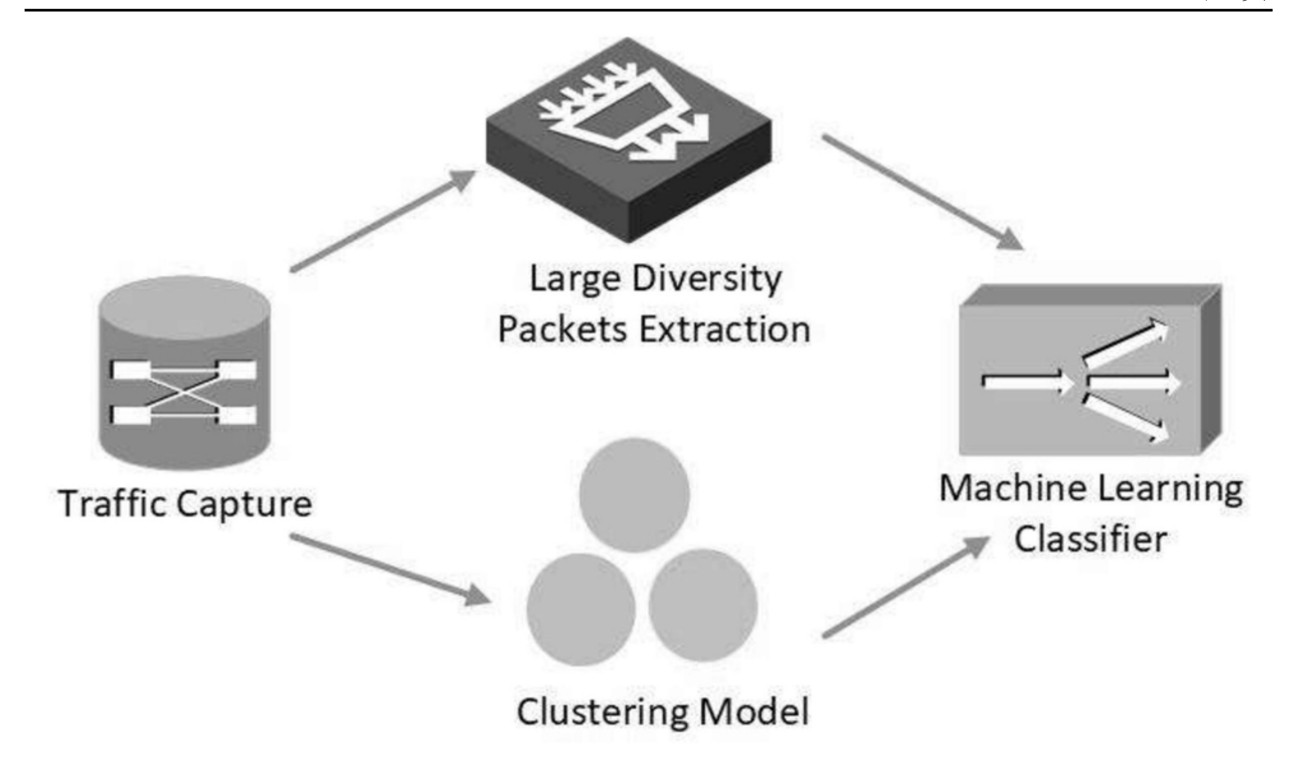

Large Diversity
Packets Extraction
Traffic Capture
Machine Learning
Classifier
Clustering Model

图1

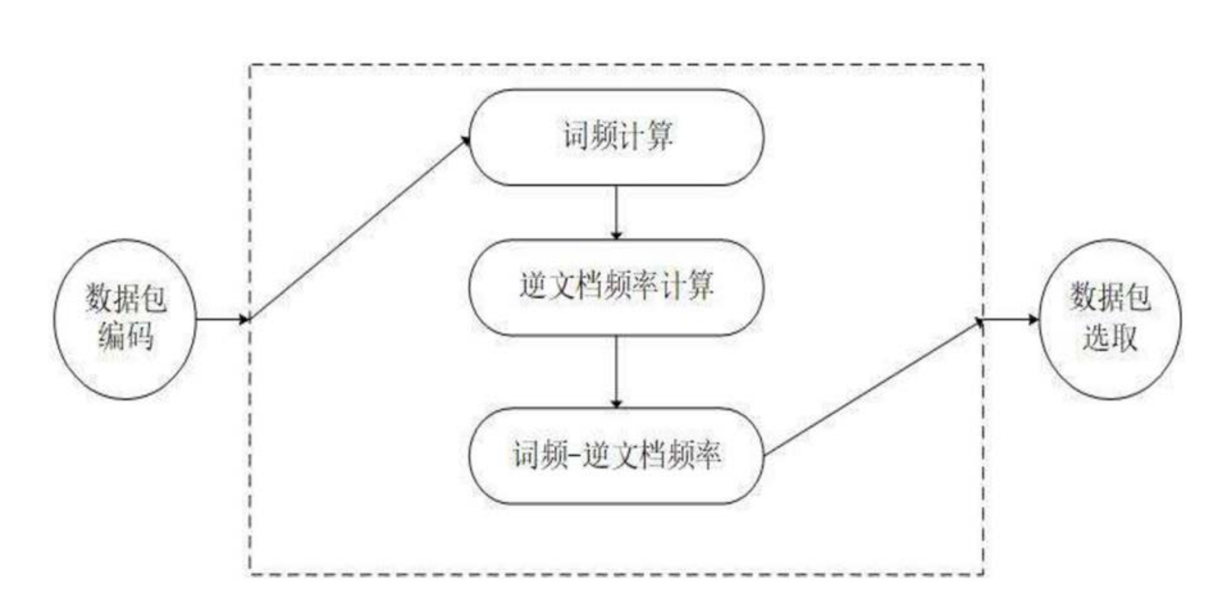

词频计算
逆文档频率计算
词频-逆文档频率
数据包
编码
数据包
选取

图2

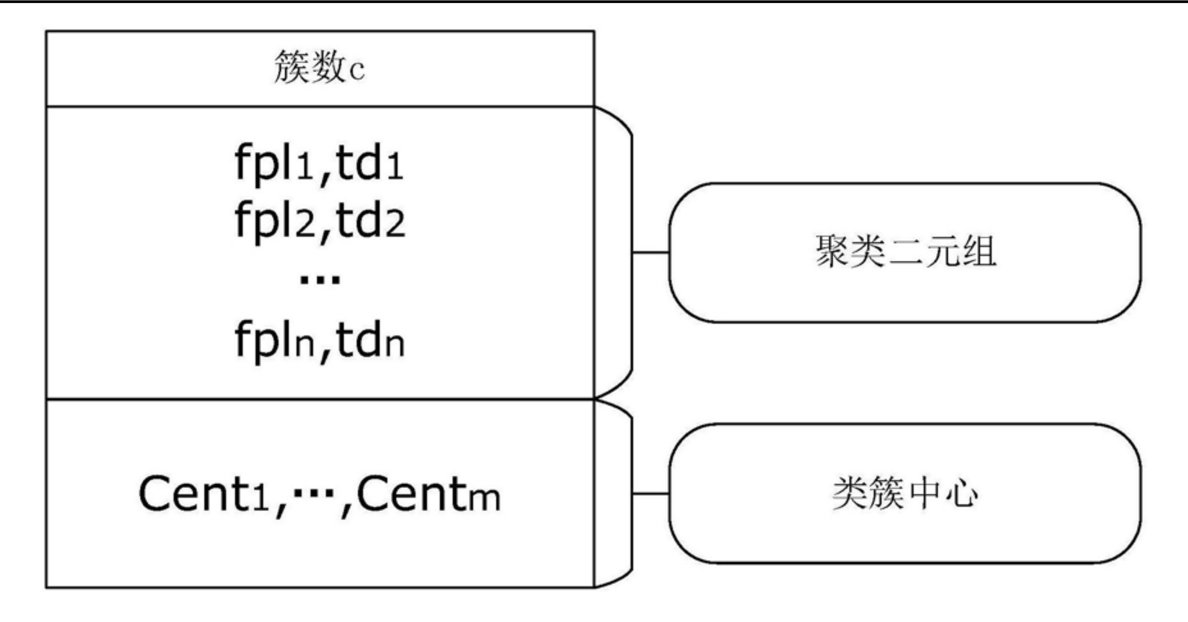


图3